Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

ROBUST BAYESIAN DESIGN AND ANALYSIS OF CLINICAL TRIALS VIA PRIOR PARTITIONING[1]

BY DANIEL J. SARGENT AND BRADLEY P. CARLIN

University of Minnesota

Unlike traditional approaches, Bayesian methods enable formal combination of expert opinion and objective information into interim and final analyses of clinical trials data. However, in cases where a broad group must be convinced by the results, a practical approach for studying and communicating the robustness of conclusions to prior specification is required. Rather than adopt the traditional method of modifying a single, initial prior and repeating the posterior calculation, in this paper we give a partial characterization of the class of priors leading to a given decision (such as stopping the trial and rejecting the null hypothesis) conditional on the observed data. We employ an interval null hypothesis based on the indifference zone approach of Freedman and Spiegelhalter, and restrict attention to priors having certain prespecified quantiles. We illustrate the application of our approach to interim monitoring using data from a recent AIDS clinical trial. We also indicate the method's usefulness in the design of future trials, creating simulation-based Bayesian analogues of the classical sample size table.

1. Introduction. Recently, Bayesian methods have seen increasing usage in the design, interim monitoring, and final analysis of clinical trials data. They allow for greatly simplified designs, due to the independence of the inference from the stopping rule, as well as more realistic sample size determination based on the full range of the experimenter's prior beliefs. Advanced Monte Carlo integration algorithms such as the Gibbs sampler enable fast and accurate computation of relevant posterior distributions, providing a more informative estimate of the treatment effect and the associated uncertainty. Moreover, Bayesian methods free the user from prespecifying the number of looks at the data or the form of an "α-spending function" (see e.g. Carlin et al., 1993). Finally, the Bayesian methodology is easily blended with formal decision-theoretic tools in settings where policymakers must do more than simply summarize a trial's results (e.g., in determining whether it is ethical to run a given trial in the first place). Thorough reviews of the use of Bayesian methodology in clinical trials are provided by Berry (1993) and Spiegelhalter, Freedman and Parmar (1994).

Despite these potential advantages, many practitioners are either skeptical of Bayesian methods or reluctant to use them. This apprehension is often

[1] Partial support for this work provided by CPCRA Contract NO1-A1-05073 from the Division of AIDS, National Institute of Allergy and Infectious Diseases. In addition, the work of the second author was supported in part by NIAID FIRST Award 1-R29-AI33466.

Key words and phrases. Indifference zone, interim analysis, stopping rule.

due to the dependence of conclusions on the particular form chosen for the prior distribution of the parameters in the model. The usual response to this problem is to repeat the analysis using a different (but still plausible) prior, and check to see if this produces a noticeable change in conclusions. Spiegelhalter et al. (1994) implement this approach for clinical trials by performing the analysis using a collection of priors chosen to reflect a broad range of opinions as to the potential benefit of the treatment. More specifically, they suggest investigating the results under a "clinical" prior, representing the (typically optimistic) prior feelings of the trial's investigators, a "skeptical" prior, reflecting the opinion of a person or regulatory agency that doubts the treatment's effectiveness, and a "noninformative" prior, a neutral position that leads to posterior summaries formally equivalent to those produced by standard maximum likelihood techniques. Agreement among the inferences drawn under all three of these priors suggests that the data are strongly informative and the precise form of prior distribution is irrelevant. Disagreement precludes a single "correct" summary of the trial, but still serves to quantify the range of plausible treatment effects and the sensitivity of the conclusions to the prior. A similar but broader approach is advocated by Greenhouse and Wasserman (1995), who compute bounds on posterior expectations over an ϵ-contaminated class of prior distributions.

An alternative to this "forward" approach to prior robustness (where one respecifies the prior and recomputes the result) is the "backward" approach of Carlin and Louis (1995). These authors start with a dataset, and then attempt to characterize the class of priors that lead to a particular conclusion (e.g., stopping the trial and deciding in favor of the treatment). Since this class can be quite large, they suggest restricting attention to "plausible" priors, such as those having a certain mean, a certain mean and variance, or certain quantiles (e.g. median, or 5^{th} and 95^{th} percentiles).

Carlin and Louis (1995) refer to this approach as *prior partitioning*, and obtain fairly specific results for some of the restricted nonparametric prior classes mentioned above. Their investigation applies to the case of a point null hypothesis and a two-sided alternative for the treatment effect θ, i.e., $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. As such, their calculations are reminiscent of those often done in studies of Bayesian robustness, an area pioneered by Edwards, Lindman, and Savage (1963). More recently, Berger and Sellke (1987) and Berger and Delampady (1987) showed that the minimum of $P(\theta = 0|x)$ over all conditional priors G for $\theta \neq \theta_0$ is attained when G places all of its mass at $\hat{\theta}$, the maximum likelihood estimate of θ. Even in this case, where G is working with the data against H_0, these authors showed that the resulting $P(\theta = 0|x)$ values are typically still larger than the corresponding two-sided p-value, suggesting that the standard frequentist approach is biased against H_0 in this case.

A somewhat more realistic approach to clinical trials would involve an interval null hypothesis $H_0 : \theta \in [\theta_L, \theta_U]$, where $[\theta_L, \theta_U]$ is some prespecified *indifference zone*, within which we are indifferent as to the use of treatment or placebo. For example, we might take $\theta_U > 0$ if there were increased costs or toxicities associated with the treatment. Here, prior partitioning would have more in common with the work of O'Hagan and Berger (1988), who obtain bounds on the posterior probability content of each of a collection of intervals which form the support of a univariate parameter, under the restriction that the prior probability assignment to these intervals is in a certain sense unimodal.

The remainder of this paper is organized as follows. In Section 2 we review the specifics of prior partitioning for clinical trial monitoring under point null hypotheses, and then extend the methodology to handle interval null hypotheses. The technology is applied to the interim analysis of a particular AIDS clinical trial dataset in Section 3, while Section 4 investigates its usefulness in the design of future trials of this type. Finally, Section 5 discusses our findings and suggests avenues for further research.

2. Prior partitioning. Consider first the point null testing scenario described in the previous section and investigated by Carlin and Louis (1995). Without loss of generality we set $\theta_0 = 0$. Suppose we are given an observation x having density $f(x|\theta)$, where θ is an unknown scalar treatment effect parameter. Let π represent the prior probability of H_0, and $G(\theta)$ the prior cumulative distribution function (cdf) of θ conditional on $\{\theta \neq 0\}$. Then the complete prior cdf for θ is given by $F(\theta) = \pi I_{[0,\infty)}(\theta) + (1-\pi)G(\theta)$, where I_S is the indicator function of the set S. The posterior probability of the null hypothesis is therefore

$$P_G(\theta = 0|x) = \frac{\pi f(x|0)}{\pi f(x|0) + (1-\pi)\int f(x|\theta)dG(\theta)} .$$

For a given prior distribution G and some $p \in (0,1)$, we will stop the experiment and reject the null hypothesis if $P_G(\theta = 0|x) \leq p$. Elementary calculations show that characterizing this class of priors $\{G\}$ is equivalent to characterizing the set $\mathcal{H}_c$, defined as

$$\mathcal{H}_c = \left\{ G : \int f(x|\theta)dG(\theta) \geq c = \frac{1-p}{p}\frac{\pi}{1-\pi}f(x|0) \right\} .$$

Carlin and Louis (1995) establish results regarding the features of $\mathcal{H}_c$, and then use these results to obtain sufficient conditions for $\mathcal{H}_c$ to be nonempty for classes of priors satisfying various moment and percentile restrictions.

Turning to the interval null hypotheses $H_0 : \theta \in [\theta_L, \theta_U]$ and $H_1 : \theta \notin [\theta_L, \theta_U]$, let π again be the prior probability of H_0, and let $G(\theta)$ now correspond to the prior cdf of θ given $\theta \notin [\theta_L, \theta_U]$. Making the simplifying assumption of a uniform prior over the indifference zone, the complete prior density function for θ may be written as

$$p(\theta) = \frac{\pi}{\theta_U - \theta_L} I_{[\theta_L, \theta_U]}(\theta) + (1 - \pi) g(\theta) .$$

Using Bayes rule, the posterior probability of H_0 is then

$$(1)\ P_G(\theta \in [\theta_L, \theta_U] | x) = \frac{\int_{\theta_L}^{\theta_U} f(x|\theta) \left[\frac{\pi}{\theta_U - \theta_L} I_{[\theta_L, \theta_U]}(\theta) + (1-\pi) g(\theta)\right] d\theta}{\int f(x|u) \left[\frac{\pi}{\theta_U - \theta_L} I_{[\theta_L, \theta_U]}(u) + (1-\pi) g(u)\right] du} .$$

We want to describe the priors G that lead to rejecting H_0, i.e., those for which (1) is less than or equal to some prespecified probability p. Since $g(\theta)$ has no support on the interval $[\theta_L, \theta_U]$, this is equivalent to describing the set

$$(2)\ \mathcal{H}_c = \left\{ G : \int f(x|\theta) dG(\theta) \geq c = \frac{1-p}{p} \frac{\pi}{1-\pi} \frac{1}{\theta_U - \theta_L} \int_{\theta_L}^{\theta_U} f(x|\theta) d\theta \right\} .$$

To restrict the class of candidate G's somewhat, suppose we consider only those for which $P_G(\theta \leq \xi_L) = a_L$ and $P_G(\theta > \xi_U) = a_U$ for some fixed ξ_L and ξ_U, where a_L and a_U lie in the unit simplex. That is, the only restriction on our prior cdf G is that it must pass through the points (ξ_L, a_L) and $(\xi_U, 1 - a_u)$. We further assume that $\max(\xi_L, \theta_L) \leq \min(\xi_U, \theta_U)$, and that $f(x|\theta)$ is a unimodal function of θ for fixed x that vanishes in both tails. Due to the asymptotic normality of the observed likelihood function, this final assumption will be at least approximately true for large datasets.

We can then derive $\sup_G \int f(x|\theta) dG(\theta)$ and $\inf_G \int f(x|\theta) dG(\theta)$, where the sup and inf are over the restricted class of G's described in the preceding paragraph. (These expressions are fairly complicated for the general case, and hence are relegated to the appendix.) Since $\mathcal{H}_c$ is empty if the sup does not exceed c, the supremum expression can be used to determine whether there are any G satisfying equation (2), i.e., whether any priors G exist that enable stopping to reject the null hypothesis. Similarly, the infimum expression may be useful in determining whether any G enable stopping to reject the alternative hypothesis, H_1. Note that in either case, we may view as fixed the (ξ_L, ξ_U) pair, the (a_L, a_U) pair, or both. For example, suppose we seek the G consistent with rejection of H_0 for a fixed (ξ_L, ξ_U) pair. Given values of $f(x|\xi_L)$, $f(x|\xi_U)$, $f(x|\hat{\theta})$, and $\int_{\theta_L}^{\theta_U} f(x|\theta) d\theta$, where $\hat{\theta}$ is the maximum likelihood estimate of θ, this amounts to determining the (a_L, a_U) pairs compatible with (2). The following two sections illustrate these ideas in the context of interim monitoring and experimental design, respectively.

3. Application to interim analysis. We apply the methodology of the preceding section to a clinical trial dataset originally analyzed by Jacobson et al. (1994), and also considered by Carlin et al. (1993) and Carlin and Louis (1995). The data are from a double-blind randomized trial comparing the drug pyrimethamine with placebo for preventing toxoplasmic encephalitis (TE), a major cause of morbidity and mortality among people with AIDS. For the likelihood, we adopt the proportional hazards model where the response variable is the time from randomization until development of TE or death. We use a Cox model having two covariates for each patient: baseline CD4 cell count, and a treatment effect indicator (1 for active drug, 0 for placebo). Denoting the parameters corresponding to these two covariates as β and θ, respectively, we obtain a marginal partial likelihood for θ by numerically integrating β out of the Cox partial likelihood. Following Section 2, we denote this marginal likelihood as $f(x|\theta)$.

3.1. Searching over (a_L, a_U) for fixed (ξ_L, ξ_U). Suppose that we wish to find pairs (a_L, a_U) for which $\mathcal{H}_c$ is non-empty given a fixed region (ξ_L, ξ_U). Since the problem formulation provides us with the indifference zone (θ_L, θ_U), it seems most natural to set $\xi_L = \theta_L$ and $\xi_U = \theta_U$. Note that this automatically implies that $a_L + a_U = 1$, since G has no support over the indifference zone.

In our Cox model, negative values of θ correspond to an efficacious treatment, so for this illustration we take $\theta_U = 0$ and $\theta_L = \log(.75) = -.288$. That is, any positive value for θ favors the placebo. However, due to its increased cost and toxicity the treatment will be preferred only for θ values smaller than $\log(.75)$, i.e., only if it reduces the placebo hazard rate by at least 25%. At the trial's fourth monitoring point, by which time $n = 60$ persons have been observed to die or contract TE, we obtain the values $f(x|\xi_L) = .02$, $f(x|\xi_U) = .18$, and $f(x|\hat{\theta}) = 1.28$ where $\hat{\theta} = .62$. Since $\hat{\theta} > \xi_U$, from the fifth row of appendix expression (8) we have that the (a_L, a_U) pairs that satisfy the condition $\sup_G \int f(x|\theta)dG(\theta) > c$ are

$$\left\{(a_L, a_U) : a_L f(x|\xi_L) + a_U f(x|\hat{\theta}) \geq \frac{1-p}{p}\frac{\pi}{1-\pi}\frac{1}{\xi_U - \xi_L}\int_{\xi_L}^{\xi_U} f(x|\theta)d\theta\right\} . \tag{3}$$

Note that under the Cox model, exact evaluation of the integral in the above expression requires numerical methods. But our dataset is large, and so $f(x|\theta)$ viewed as a function of θ is well-approximated by a normal distribution with mean $\hat{\theta}$ and standard deviation .312. Thus the integral in (3) reduces to a difference of two normal cdf values.

If we select $p = .1$ and $\pi = .25$, the inequality in (3) becomes $a_U \geq -.0156 a_L + .176$. But recall that $a_L + a_U = 1$, so the set of (a_L, a_U) pairs for which at least one prior exists that leads to the rejection of H_0 can be represented as the line segment defined by $\{(a_L, a_U) : a_L + a_U = 1, a_U \geq .163\}$. The usefulness of this fact is best seen by inverting it: There are *no* priors having $a_U < .163$ that enable rejection of the null hypothesis. This is a plausible outcome for this dataset: using a prior with very little support for positive θ values, the data (which *do* support these values) are not yet sufficiently convincing, and the trial must be continued. For all other a_U values, stopping to reject H_0 is at least possible.

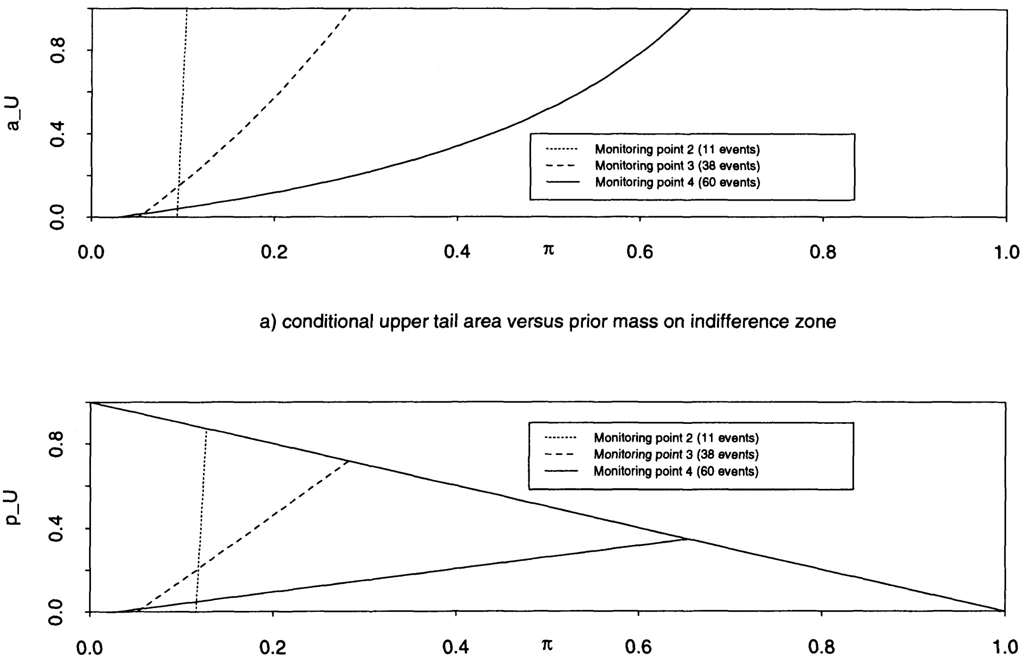
To expand the scope of our analysis, we might replace a_L by $1 - a_U$ in (3), and solve for a_U as a function of π. This produces the inequality

$$a_U \geq .536 \frac{\pi}{1-\pi} - .016 \ . \tag{4}$$

Combining (4) with the constraints $0 \leq a_U \leq 1$ and $0 \leq \pi \leq 1$ produces the regions determined by the solid curve in Figure 1(a). For (π, a_U) combinations above and to the left of this curve, there exist priors consistent with that combination which permit stopping to reject H_0, while for combinations below and to the right of the curve, no such priors exist. That is, no prior with a (π, a_U) combination lying below the curve would lead to rejection of the null hypothesis.

Figure 1(a) also shows the boundaries obtained in the same manner as formula (4) for monitoring points two ($n = 11$ events, $\hat{\theta} = .02$) and three ($n = 38$ events, $\hat{\theta} = .49$). In the first case, all but the most extreme priors (those having $\pi < .104$) preclude stopping to reject H_0. As the number of observed events n and $\hat{\theta}$ increase over time, the potential stopping regions lying to the left of the curves also increase in size. This emerging superiority of the placebo was not anticipated by any of the five subject area experts consulted for this trial (Chaloner et al., 1993), so our prior partitioning analysis seems an especially useful complement to traditional Bayesian analysis in this problem.

To ease the interpretability of our results, we might replace the conditional upper tail probability a_U with the corresponding unconditional probability, $p_U \equiv a_U(1 - \pi)$. This converts (4) into a linear inequality, but with the added constraint that $p_U + \pi \leq 1$. Figure 1(b) plots the (π, p_U) pairs and their status relative to stopping and rejecting H_0 for the same three monitoring points shown in Figure 1(a). This plot may be easier for a clinician to interpret, since it avoids the notion of a_U, a probability that is conditional on the null hypothesis being false. Again, no prior corresponding to a region to the right of a boundary enables stopping to reject the null hypothesis at this monitoring point. Hence if a clinician had a relatively small p_U (low credence on large values for θ that suggest superiority of the placebo) coupled with a relatively large π (high credence on θ values in the indifference zone), H_0 could not be rejected even at the fourth monitoring point. But note also that no point below the solid line has $p_U = .5$, which for this problem



a) conditional upper tail area versus prior mass on indifference zone

b) unconditional upper tail area versus prior mass on indifference zone

Figure 1: *Prior tail area regions at three monitoring points, TE trial data. For combinations to the left of each curve, priors exist that permit stopping to reject H_0.*

defines the class of priors that are "skeptical" (symmetric about 0). Hence for each potential value of π, there is at least one skeptical prior that would be convinced as to the placebo's superiority by monitoring point four.

From equation (4), rejection of H_0 is always possible for $\pi < .029$, while it is never possible for $\pi > .655$. These are sensible results, since low prior weight on the null should encourage rejection, while sufficiently high prior weight on the null should prevent it. For comparison, the analysis of Carlin and Louis (1995) based on the point null prior $H_0 : \theta = 0$ produces the π cutoffs .014 and .410, respectively. These numbers are also plausible, since the maximum likelihood estimate (MLE) $\hat{\theta} = .62 > 0$, and so it should be more difficult to reject a null hypothesis that concentrates all its mass at 0 than one which distributes the same mass uniformly across an interval with upper boundary 0. Of course, in practice one might well use a smaller value of π with the point null than with the interval null, so the two approaches suggest similar stopping patterns.

One of the advantages of a Bayesian approach is that it enables formal rejection of the alternative hypothesis, should the data support such a move. We would take such action if $P_G(\theta \notin [\theta_L, \theta_U]|x) \leq p$, or equivalently if

$$(5) \qquad \int f(x|\theta)dG(\theta) \leq c' = \frac{p}{1-p}\frac{\pi}{1-\pi}\frac{1}{\theta_U-\theta_L}\int_{\theta_L}^{\theta_U} f(x|\theta)d\theta \, .$$

Hence we are now interested in appendix formula (7), which gives the infimum of $\int f(x|\theta)dG(\theta)$. But since $(1 - a_L - a_U) = 0$, this infimum will also be 0. Thus there is at least one G such that rejection of the alternative is possible for *all* π and a_U. Hence we can say little about the evidence against the alternative in the case where $\xi_L = \theta_L$ and $\xi_U = \theta_U$.

As a possible remedy for this poor resolution, we might instead fix (ξ_L, ξ_U) such that $\xi_L < \theta_L < \theta_U < \xi_U$. In this case we do not have the constraint that $a_L + a_U = 1$, enabling the same sort of analysis as above, but with a nondegenerate solution to equation (7). Suppose we take $\xi_L = \theta_L - .5 = -.788$ and $\xi_U = \theta_U + .5 = .5$. Since we will no longer have that $a_L = 1 - a_U$, we again fix π and seek (a_L, a_U) pairs that lead to stopping. Inserting the infimum from equation (7) into inequality (5), we obtain that stopping to reject the alternative is possible if

$$(6) \qquad a_U \geq (1 - c'/M) - a_L \, ,$$

where $M = \min(f(x|\xi_L), f(x|\xi_U))$. This means that (a_L, a_U) pairs that enable stopping will lie above and to the right of a line having slope -1.

While this approach may prove useful for a given dataset, there remains a problem that limits its application in general. For many families f (including the normal), when $\hat{\theta}$ is large it will often be the case that $c' > M$, and so equation (6) will be vacuous. Then for every (a_L, a_U) pair there will be at least one prior that enables rejection of H_1, even though the data support this hypothesis. The explanation for this counterintuitive behavior is that the prior corresponding to the infimum in (7) places all its mass at only 3 support points, none of which are supported by the data. In Section 5 we suggest several modifications to this approach to improve its utility.

3.2. Searching over (ξ_L, ξ_U) *for fixed* (a_L, a_U). As an alternative method of partitioning the prior space, we might reverse the procedure of the previous subsection and search over prior quantiles (ξ_L, ξ_U) corresponding to prespecified tail areas (a_L, a_U). Here, a_L and a_U would typically be fixed at small values, say, $a_L = a_U = .025$. Then it would likely be the case that $\xi_L << \theta_L < \theta_U << \xi_U$, for otherwise, with so much prior mass concentrated near the indifference zone, we would have ethical doubts regarding whether the trial should be conducted at all.

The question arising in this case then is for this fixed pair (a_L, a_U), do there exist any priors that lead to rejection of the null hypothesis? The answer is again governed by the inequality in (2). As in the previous subsection it is reasonable to look for the supremum of $\int f(x|\theta)dG(\theta)$ over G, but such a solution will be complicated in general, for as ξ_L and ξ_U range across their set of possible values, the supremum will range over all five cases listed in appendix equation (8).

In specific applications, however, θ_L, θ_U, and $\hat{\theta}$ will be known, so with the assumption that $\max(\xi_L, \theta_L) \leq \min(\xi_U, \theta_U)$, many of these five cases will become impossible. For our TE dataset with p, π, a_L, and a_U as previously specified, solving for the supremum at the fourth monitoring point involves the cases on lines 4 and 5 of appendix equation (8). Line 4 yields no constraints on ξ_L as long as $\xi_U \geq \hat{\theta} = .62$, while line 5 produces the equation $f(x|\xi_U) \geq .204 - .024 f(x|\min(\xi_L, \theta_L))$ when $\xi_U < .62$. These equations imply that there exist priors that will allow rejection of the null hypothesis depending on the values of ξ_L and ξ_U as follows:

1. For $\xi_L \to -\infty$, we must have $\xi_U \geq .0219$
2. For $\xi_L \geq -.288$, we must have $\xi_U \geq .0215$
3. For $\infty < \xi_L < -.288$, we must have $f(x|\xi_U) \geq .204 - .024 f(x|\xi_L)$, that is, $\xi_U \geq \omega$ where ω ranges from .0219 to .0215.

So for this dataset, any prior having 97.5^{th} percentile smaller than .0215 precludes stopping to reject the null hypothesis; the value of such a prior's 2.5^{th} percentile is virtually arbitrary in this decision.

4. Application to design. Sample size determination is one of the most difficult problems faced by the designers of a clinical trial. The usual approach is to prepare a "sample size table" having a range of power levels (say, 80, 90, and 95 percent) as its column headings, and likely treatment effect magnitudes as its row headings. The table entries then give the sample sizes required to detect the various treatment effects at the various power levels. While these calculations are exact, they must be used cautiously as they involve guesses for important characteristics of the errors involved in the measurement process.

The necessity of subjectively assessing the uncertainty associated with the true treatment effect makes trial design an inherently Bayesian procedure. Indeed, such methods have a longer history of use in the design of clinical trials than in their monitoring or final analysis (see e.g. Freedman and Spiegelhalter, 1983). In this section we show how prior partitioning can be used in a trial's design stage by determining which combinations of true treatment effect and trial sample size are likely to lead to definite conclusions (i.e., conclusions that hold for a broad range of prior distributions on the treatment effect). Consider again a treatment effect parameter θ having MLE $\hat{\theta}$, which for convenience we assume has a distribution that can be reasonably well approximated as $N(\theta, \sigma^2)$. For a postulated (θ, σ^2) pair, we could draw independent samples $\{\hat{\theta}_k,\ k = 1, \ldots, K\}$, and apply our prior partitioning approach using each simulated MLE. A suitable numerical or graphical summary of the results could inform as to whether successful stopping of the trial was likely for this (θ, σ^2) pair.

As a concrete illustration, consider again the TE trial setting investigated in Subsection 3.1, where we search over (a_L, a_U) pairs having $\xi_U = \theta_U = 0$ and $\xi_L = \theta_L = \log(.75) = -.288$. Appendix expression (8) then determines three possible cases depending on whether the generated $\hat{\theta}_k$ falls in $(-\infty, \xi_L]$, $(\xi_L, \xi_U]$, or (ξ_U, ∞). Recalling that $a_L = 1 - a_U$ in this case, and again fixing $p = .1$ and $\pi = .25$, each replication produces a (possibly empty) interval of a_U values, for each of which there exists at least one prior that leads to stopping and rejecting H_0.

Table 1: *Summaries of a_U intervals for which there exists a prior that enables stopping to reject H_0; $\theta_L = \xi_L = -.288$, $\theta_U = \xi_U = 0$, $p = 0.1$, and $\pi = 0.25$.*

	$n = 250$		$n = 1000$		$n = 2000$	
true θ	prop nonempty	avg length	prop nonempty	avg length	prop nonempty	avg length
-0.144	0.14	0.06	0.12	0.06	0.11	0.04
0.0	0.16	0.08	0.18	0.09	0.26	0.14
0.5	0.48	0.31	0.91	0.76	1.00	0.96
1.0	0.86	0.71	1.00	1.00	1.00	1.00

Table 1 summarizes our results in a way intended to mimic a traditional sample size table. As usual, the row headings are potential values of the true treatment effect θ, beginning with $\theta = -.144$, the midpoint of the indifference zone. But the column headings now correspond to possible sample sizes n in each treatment arm, which for the purpose of this illustration we convert to a likelihood variance using the formula $\sigma^2 = 50/n$. (This would be the appropriate formula for $Var(\hat{\theta})$ if instead of the Cox model we simply assumed the observations X_i in each group to be independent normal variables with variance 25, and we took $\hat{\theta} = \bar{X}_{drug} - \bar{X}_{placebo}$.) For each (θ, n) combination, the table provides the proportion of nonempty a_U intervals and the average length of these intervals over $K = 1000$ simulated replications. As we would expect, using either summary measure, the likelihood of the trial stopping and rejecting the null hypothesis increases as θ moves away

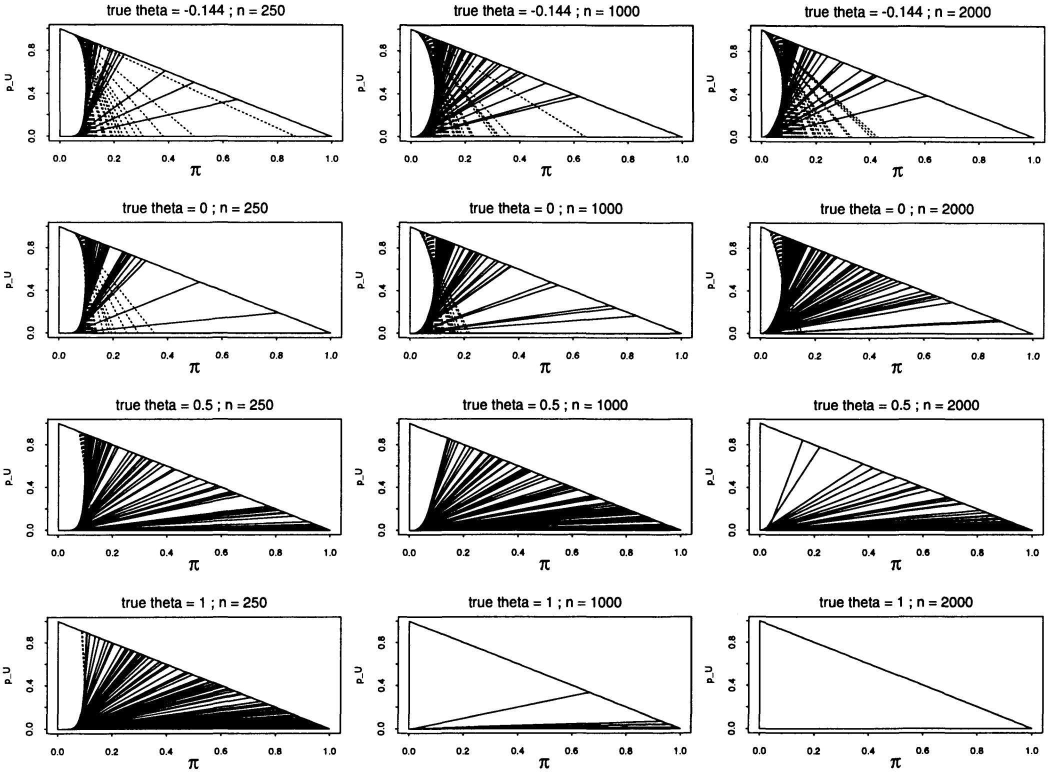


Figure 2: *Simulated regions for stopping and rejecting H_0, where $\theta_L = \xi_L = -0.288$, $\xi_U = \theta_U = 0$, and $p = 0.1$. For (π, p_U) combinations above and to the left of the solid lines and below and to the left of the dashed lines, priors exist that permit stopping to reject H_0.*

from the indifference zone. Larger sample sizes lead to a similar increase for θ values not in the center of the indifference zone. While this table could be reduced to a single pair of numbers by placing hyperpriors on θ and n, this sort of tabular display is likely to be more popular among practitioners.

In addition to these a_U interval summaries for fixed π, we might also consider plots of p_U versus π, as in Figure 1(b). Figure 2 plots the boundaries obtained over $K = 100$ replications from each possible combination of the four sample sizes and four true treatment effects investigated in Table 1. That is, there are 100 lines plotted in each of the 16 graphs, though many of these lines are not visible in the lower-right graphs because they coincide with the π-axis. The (π, p_U) pairs for which there exists a prior that permits stopping to reject H_0 are located above and to the left of the boundaries having positive slope (plotted as solid lines), but below and to the left of those boundaries having negative slope (plotted as dashed lines). Thus for a given graph, the intersection of these regions gives the (π, p_U) pairs for which stopping to reject H_0 was possible in every simulated replication. In most of the 16 cases shown this is only a small convex area on the left side of the graph, but it is larger for larger true treatment effects and sample sizes (for $\theta = 1$ and $n = 2000$, it is the entire region). Conversely, the intersection of the complements of the regions describes the (π, p_U) pairs for which *no* prior enabled stopping to reject H_0 in any of the replications. Note that this is a fairly large region on the right side of the graph in which $\theta = -.144$ and $n = 2000$, a situation where we are very likely to have strong evidence of no treatment effect.

While the purpose of Figure 2 is primarily to convey a rough visual impression of the variability in the strength of evidence likely to be obtained from various (θ, n) combinations, the intersection regions described in the previous paragraph could be interpreted more formally in the context of randomization tests (Barnard, 1963). That is, since we have used $K = 100$, these regions are essentially areas wherein we are just over 99% confident as to the outcome (ability to reject H_0 for some prior or for no prior, respectively). Of course different sets of random numbers will lead to different confidence regions, but these differences should be small in the cases wherein we would actually contemplate running a trial (i.e., those like the $(\theta = 1, n = 1000)$ and $(\theta = 1, n = 2000)$ cases in Figure 2, where virtually the entire region permits stopping to reject the null).

5. Discussion. While our methods do not constitute a replacement for formal Bayesian methods based on careful prior elicitation, they do offer a useful way of "scoping" the prior class and identifying those that are in sharp conflict with the data. We anticipate their forming a useful first step in a Bayesian analysis, to determine whether there exists a prior that would lead to rejection of the null (or the alternative) given some very vaguely specified prior beliefs, as quantified by a particular combination of π, ξ_L, ξ_U, a_L, and a_U. An unambiguous answer may be helpful to a data safety and monitoring board in determining that a clinical trial should be stopped. An ambiguous answer requires more careful prior specification, a continuation of the trial to accumulate more data, or both.

In the context of experimental design, the results in this paper are based on approximate normality. This is a fairly common assumption in practice, and with good reason: besides the asymptotic normality of the MLE, Tsiatis (1981) shows that $4L/n \mathrel{\dot\sim} N(\delta, 4/n)$ in the vicinity of the null hypothesis, where L is the usual log-rank test statistic and δ is the log-relative hazard between the two treatment groups. Still, the assumption of normality may be too restrictive in practice. However, note that we assumed normality only for computational convenience; in principle, prior partitioning applies equally well to nonnormal likelihoods, including proportional hazards models. Another worthwhile extension would be to consider multivariate θ, which would be required to analyze multi-arm trials or study effectiveness and toxicity simultaneously.

The priors corresponding to the boundaries in Figures 1 and 2 are "extreme," in the sense that they consist of point masses placed at only three values on the real line. Besides being unrealistic as representations of any individual's beliefs, these extremal priors lead to difficulties in using equation (5) to determine prior partitions for rejecting H_1 (i.e., accepting H_0), as described near the end of Subsection 3.1. A natural solution to this problem would be to further restrict the class of candidate priors, screening out those which are too extreme. For example, we could limit our attention to ϵ-contamination priors. Alternatively, we might consider only continuous priors for θ that satisfy some smoothness condition(s), imposed by constraining the derivative(s) of $g(\theta)$. Most drastically, we could restrict attention to priors having a certain parametric form (e.g., the normal family). While this would likely lead to fairly precise results at little computational expense, it would also eliminate a large number of plausible priors (e.g., the Student's t family). We hope to report on many of these approaches in a subsequent paper.

Acknowledgements. The authors thank Profs. Jim Hodges, Tom Louis, and Larry Wasserman for several suggestions that greatly improved the manuscript.

APPENDIX

As described in Section 2, suppose we have specified an indifference zone (θ_L, θ_U), and we restrict our attention to conditional priors G satisfying $P_G(\theta \leq \xi_L) = a_L$ and $P_G(\theta > \xi_U) = a_U$. Assuming that $\max(\xi_L, \theta_L) \leq \min(\xi_U, \theta_U)$, and that the likelihood $f(x|\theta)$ is unimodal and approaches 0 for $x \to \pm\infty$, we have that

$$\inf_G \int f(x|\theta)dG(\theta) = (1-a_L-a_U)\times\min\begin{pmatrix} \delta(\theta_L - \xi_L)f(x|\xi_L), \\ \delta(\xi_U - \theta_U)f(x|\xi_U), \\ \delta(\theta_L - \xi_L)\bar{\delta}(\xi_U - \theta_U)f(x|\theta_L), \\ \bar{\delta}(\theta_L - \xi_L)\delta(\xi_U - \theta_U)f(x|\theta_U) \end{pmatrix}, \tag{7}$$

where $\delta(x) = I_{(0,\infty)}(x)$, and $\bar{\delta}(x) = 1 - \delta(x) = I_{(-\infty,0]}(x)$. Further, we have that

$$
\begin{aligned}
\sup_G \int f(x|\theta)dG(\theta) = {} & a_L \left\{\begin{matrix} f(x|\hat{\theta}) \\ f(x|\max(\xi_L,\theta_L)) \\ f(x|\max(\xi_L,\theta_L)) \\ f(x|\max(\xi_L,\theta_L)) \\ f(x|\max(\xi_L,\theta_L)) \end{matrix}\right\} + a_U \left\{\begin{matrix} f(x|\max(\xi_U,\theta_U)) \\ f(x|\max(\xi_U,\theta_U)) \\ f(x|\max(\xi_U,\theta_U)) \\ f(x|\max(\xi_U,\theta_U)) \\ f(x|\hat{\theta}) \end{matrix}\right\} \\
& +(1-a_L-a_U)\left[\delta(\theta_L-\xi_L)\bar{\delta}(\xi_U-\theta_U)\left\{\begin{matrix} f(x|\xi_L) \\ f(x|\hat{\theta}) \\ f(x|\theta_L) \\ f(x|\theta_L) \\ f(x|\theta_L) \end{matrix}\right\}\right. \\
& \qquad +\bar{\delta}(\theta_L-\xi_L)\delta(\xi_U-\theta_U)\left\{\begin{matrix} f(x|\theta_U) \\ f(x|\theta_U) \\ f(x|\theta_U) \\ f(x|\hat{\theta}) \\ f(x|\xi_U) \end{matrix}\right\} \\
& \qquad \left. +\delta(\theta_L-\xi_L)\delta(\xi_U-\theta_U)\left\{\begin{matrix} f(x|\xi_L) \\ f(x|\hat{\theta}) \\ \max(f(x|\theta_L), f(x|\theta_U)) \\ f(x|\hat{\theta}) \\ f(x|\xi_U) \end{matrix}\right\}\right] ,
\end{aligned}
\tag{8}
$$

where $\hat{\theta}$ is the MLE of θ, and the five options in the braces are for $\hat{\theta}$ falling in the intervals $(-\infty, \min(\xi_L,\theta_L)]$, $(\min(\xi_L,\theta_L), \max(\xi_L,\theta_L)]$, $(\max(\xi_L,\theta_L), \min(\xi_U,\theta_U)]$, $(\min(\xi_U,\theta_U), \max(\xi_U,\theta_U)]$, and $(\max(\xi_U,\theta_U), \infty)$, respectively. These two expressions are easy, if somewhat tedious, to obtain. Broadly speaking, equation (7) arises from placing the prior mass of G as far from the MLE as allowed by the percentile constraints, while for equation (8) we place the prior mass as close to the MLE as possible. Notice that at most one of the three terms within the brackets in equation (8) will be present for any given ordering of ξ_L, ξ_U, θ_L, and θ_U. In particular, if $\xi_L \geq \theta_L$ and $\xi_U \leq \theta_U$ then $(1-a_L-a_U) = 0$, and so all three of these terms are irrelevant.

REFERENCES

Barnard, G.A. (1963). Comment on "The spectral analysis of point processes," by M.S. Bartlett. *J. Roy. Statist. Soc. Ser. B* **25** 294.

Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science* **2** 317–352.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion). *J. Amer. Statist. Assoc.* **82** 112–122.

Berry, D.A. (1993). A case for Bayesianism in clinical trials (with discussion). *Statistics in Medicine* **12** 1377–1404.

CARLIN, B.P., CHALONER, K., CHURCH, T., LOUIS, T.A. and MATTS, J. (1993). Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *The Statistician* **42** 355–367.

CARLIN, B.P. and LOUIS, T.A. (1996). Identifying prior distributions that produce specific decisions, with application to monitoring clinical trials. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, K. Chaloner, and J. Geweke. Wiley, New York, 493–503.

CHALONER, K., CHURCH, T., LOUIS, T.A. and MATTS, J. (1993). Graphical elicitation of a prior distribution for a clinical trial. *The Statistician* **42** 341–353.

EDWARDS, W., LINDMAN, H., and SAVAGE, L.J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev.* **70** 193–242.

FREEDMAN, L.S. and SPIEGELHALTER, D.J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* **32** 153–160.

GREENHOUSE, J.B. and WASSERMAN, L.A. (1995) Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine* **14** 1379–1391.

JACOBSON, M.A., BESCH, C.L., CHILD, C., HAFNER, R., MATTS, J.P., MUTH, K., WENTWORTH, D.N., NEATON, J.D., ABRAMS, D., RIMLAND, D., PEREZ, G., GRANT, I.H., SARAVOLATZ, L.D., BROWN, L.S., DEYTON, L., AND THE TERRY BEIRN COMMUNITY PROGRAMS FOR CLINICAL RESEARCH ON AIDS (1994). Primary prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced human immunodeficiency virus disease: Results of a randomized trial. *J. Infectious Diseases* **169** 384–394.

O'HAGAN, A. and BERGER, J.O. (1988). Ranges of posterior probabilities for quasiunimodal priors with specified quantiles. *J. Amer. Statist. Assoc.* **83** 503–508.

SPIEGELHALTER, D.J., FREEDMAN, L.S. and PARMAR, M.K.B. (1994). Bayesian approaches to randomised trials (with discussion). *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

TSIATIS, A.A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **68** 311–315.

DIVISION OF BIOSTATISTICS
SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA
BOX 303 MAYO BUILDING
MINNEAPOLIS, MINNESOTA 55455

Robust Design and Analysis of Clinical Trials via Prior Partitioning

discussion by
JOSEPH B. KADANE
Carnegie Mellon University

The central idea of "prior partitioning" is as follows: given a set of decisions D, fixed data x, parameter space ω, and known likelihood $f(x \mid \theta)$, one supposes that the class of priors on ω is restricted to some subset P_0. Then the question is asked, for each $d \epsilon D$, is there a prior in P_0 such that d is optimal? If not, the suggestion is that d can be eliminated as a possible decision. In this respect, prior partitioning works a bit like admissibility. Perhaps a better name would be "posterior partitioning of prior distributions", since the partitioning depends on the observed data x.

Sargent and Carlin (SC) apply this general scheme to interval hypothesis testing of $H_0 : \theta \in [\theta_L, \theta_U]$ against the alternative $\theta \notin [\theta_L, \theta_U]$, taking the restricted class P_0 of priors to be those that put probability π uniformly on $[\theta_L, \theta_U]$, and $(1-\pi)$ on G, outside $[\theta_L, \theta_U]$ with two quantiles fixed, i.e. $G(\xi_L) = a_L$ and $G(\xi_U) = 1 - a_U$. They then ask whether there are priors G, for fixed π, ξ_U, a_L, a_U and p, such that $P_G[\theta \in [\theta_L, \theta_U] \mid x] \leq p$.

As an illustration, SC apply this idea to a Cox proportional hazards model with two covariates, baseline CD_4 count and treatment. They integrate out the parameter for CD_4 count, yielding a marginal Cox partial likelihood, which they take to be their likelihood $f(x \mid \theta)$. It seems to me that this introduces two unexamined possible sources of non-robustness: the partial likelihood, and the marginalization. In particular, I would give up some robustness in the treatment parameter to be better protected against possible non-robustness in the CD_4 count.

My major concern is whether it is useful to restrict attention to the posterior probability content of $[\theta_L, \theta_U]$. In their example, small values of θ favor the treatment, while large values favor the placebo. Consider situation 1 in which half the posterior probability outside $[\theta_L, \theta_U]$ is at $\theta_U + \epsilon$ while half is at $\theta_L - 1/\epsilon$. Then for small positive ϵ, choosing the treatment will be well advised because the loss of choosing the treatment will be dominated by the lump of probability at $\theta_U + \epsilon$, just outside the indifference zone. Now consider situation 2, with half the posterior probability outside $[\theta_L, \theta_U]$ at $\theta_L - \epsilon$ while half is at $\theta_U + 1/\epsilon$. Now it would be best to choose the placebo, for symmetric reasons. Nonetheless, the SC analysis, concentrating on the posterior probability of $[\theta_L, \theta_U]$ would evaluate these two situations identically.

SC, perhaps inadvertently, raise an issue about the appropriateness of the structure they propose, writing "Negative values of θ correspond to efficacious treatment, so... we take $\theta_U = 0$ and $\theta_L = \log(.75) = -2.88\ldots$ Due to increased cost and toxicity the treatment will be preferred only for values of θ smaller than log (.75), i.e. only if it reduces the hazard rate by at least 25%". This suggests to me that they would use the treatment if $\theta \leq -2.88$ and the placebo otherwise. Perhaps it is necessary to use a hypothesis-testing framework for the problem, i.e. a 0-1 loss, in which case it would seem that the relevant hypotheses are $H_0 \leq -2.88$ versus $H_1 : \theta > -2.88$. Perhaps a more realistic loss would penalize the use of the placebo if $\theta \leq -2.88$ in a way that increases as θ decreases, and penalize the use of the treatment if $\theta > -2.88$ in a way that likewise increases as θ increases.

There are difficulties in the interpretation of prior partitioning. If there are no priors in P_0 such that d is optimal, should I discard d? In problems with more available decisions, d might be a good compromise, and might be robust, although never optimal. If there are such priors in P_0, should I care what they look like, and whether they are in some sense reasonable?

In conclusion, I find that the strengths of this paper are its use of real data, and that it is aimed at applications to decisions real people have, in clinical monitoring and design. I believe that the principal area for further growth is that more attention should be paid to the decision aspect of the problem: what are the available decisions, what is a reasonable loss or utility structure. Finally, I believe that more attention should be paid to the robustness of the expected utility, not of the decision.

REJOINDER

Daniel J. Sargent and Bradley P. Carlin

We thank Prof. Kadane for his thoughtful comments. We agree that the name "prior partitioning" does obscure the fact that the partitions are made in light of the data, and that a more descriptive name for the technique may be in order. Indeed, the title of an early version of Carlin and Louis (1995) included the phrase, "data and decision based prior partitions," which led to our current abbreviated moniker.

Regarding the two sources of possible nonrobustness in our data example, first, a theoretical justification for our use of the Cox partial likelihood as a likelihood is given in Kalbfleisch (1978). This paper places a gamma process prior on the baseline hazard function (independent of the prior for the regression parameters), and shows that the marginal posterior density for the regression parameters approaches a form proportional to the Cox partial likelihood as this gamma process prior becomes arbitrarily diffuse.

From a more practical viewpoint, the overwhelming popularity of the Cox model among biostatistical practitioners helps to justify its use as a base model. Second, the CD4 count parameter β is only a nuisance parameter in our model, included to calibrate the likelihood for the baseline health status of the patients in our study. Since the treatment effect parameter θ forms the basis for subsequent decisions and is the only model quantity about which data monitoring board members are likely to have strong opinions, marginalizing β out of the Cox likelihood under a flat prior before beginning our investigation seems justified. Of course in principle, prior partitioning applies to multivariate parameters as well, but not without substantial complications to the analysis and graphical display of results.

We agree that there is a possible oversimplification inherent in our two-sided testing scenario. Our methods could be adapted to consider rejection of $H_L : \theta < \theta_L$ or $H_U : \theta > \theta_U$, instead of $H_0 : \theta \in [\theta_L, \theta_U]$. Fortunately, the extreme scenario Prof. Kadane describes using a small positive ϵ is rare in practice under our Cox model. Still, the value judgement he finds lurking in the indifference zone is real, and warrants a more complete investigation via formal decision-theoretic tools under appropriate loss functions. In the present paper we have deliberately concentrated solely on the probability aspect of such a model, but we hope to report tangible results on the more complete project in a subsequent paper.

Finally, the shape of the priors that allow stopping is clearly an important issue, especially given the breadth of our prior class. Carlin and Sargent (1995) consider the imposition of stronger restrictions on the prior (such as continuity, unimodality, and specific parametric and semiparametric forms), and investigate their impact on the size of the resulting prior partitions.

REFERENCES

Carlin, B.P. and Sargent, D.J. (1996). Robust Bayesian approaches for clinical trial monitoring. *Statistics in Medicine*, **15** 1093–1106.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *J. Roy. Statist. Soc., Ser. B*, **40**, 214–221.